SUMMARY OF THE INVENTION

The present invention provides a method and system for inputting Chinese characters into a computer. The invention improves the ease of use as well as efficiency of inputting Chinese characters over the prior art. Ease of use and efficiency are inherently conflicting goals in Chinese character input systems.

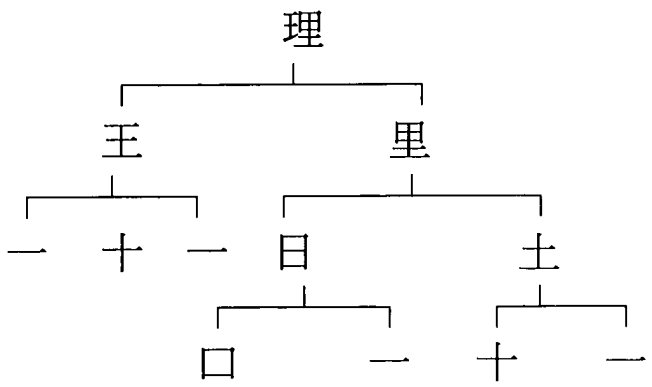
According to a first aspect of the invention, some of the 200+ components (also called radicals in the literature) used to construct Chinese characters is assigned representation by one of the letters in the English alphabet. This set of selected components is sufficient to construct any Chinese character of interest. Each Chinese character of interest to the present invention is assigned an "encoding", being a text string in the English language, with each letter of the string corresponding to the Chinese character component as defined by the present invention. This is standard practice in the prior art. In the prior art, the input systems match a given text string against the set of encodings (the library) letter for letter. An input string that matches one in the library selects the Chinese character associated with that encoding. This technique requires the user to accurately memorize the exact encoding assigned to every Chinese character, a monumental task prone to error, confusion, and forgetting from disuse. The present invention uses a novel technique in order to reduce the amount of memorization required of the user.

In a first aspect of the present method, a novel technique is used to encode Chinese characters. In the prior art, each Chinese character is encoded as a string of English letters. This string is then compared to user input in order to find a match. In the present method, each Chinese character is not encoded as a string of letters, but as a data graph. This is a technique described in the Finite State Automata field of Computer Science.

According to the theories of Finite State Automata (FSA), a Chinese character can be described by a Non-deterministic Finite State Automata (NFA). An NFA is a structure which has multiple representations in multiple parts of the structure. A Chinese character is generally composed of other simpler Chinese characters. These simpler characters are in turn composed of even simpler

characters, and so on, until finally indivisible strokes. This type of structure fits the definition of an NFA and all the techniques developed for NFA analysis can be applied. In the prior art, each Chinese character is reduced to a string, resulting in a loss of the inherent hierarchical structure of the character. Describing a Chinese character as an NFA preserves the inherent structure and has useful benefits.

For example, the character "理" can be represented by the following graph:



The interpretation of this graph is that the character "理" can be described by multiple sequences of components:

1. 王、里
2. 一十一里
3. 王日土
4. 王口一土
5. 王口一十一
6. 王日十一
7. 一十一日土
8. 一十一口一土
9. 一十一口一十一
10. 一十一日十一

The multiple descriptions of this character are a result of the character's inherent hierarchical structure, as depicted in the graph. Each distinct description

represents a unique path of traversal through the graph. Graphs like these are typically used to describe NFA's.

The benefit to the user is a reduction in the amount of memorization required in Chinese data entry. In every graph, the leaf nodes are one of a few fundamental strokes. Therefore, a beginner user only needs to memorize these few fundamental strokes and can enter any character using just these strokes, in essence traversing the bottom level of the graphs. As the user gains experience and learns more high level components, he will gradually ascend to higher level paths throught the same graph, resulting in fewer components used in describing the same character, thus increasing typing speed.

As mentioned earlier, once characters are represented as NFA's, the techniqes of Finite State Automata theory can be applied in processing the NFA's. In particular, claim 1 describes a technique wherein a whole sub-branch in a graph can be matched to a single user input symbol by equating the symbol to a string of symbols which are the flattened contents of the sub-branch. This a technique commonly known as reducing an NFA to a DFA (Deterministic Finite State Automata).

In a second aspect of the present method, a "partial match" algorithm is used to further increase the intelligence of the encoding comparison operation. Whereas explicit "wildcard" letters could be supplied by the user in an encoding, an "implicit" wildcard is automatically created by the present invention whenever a given input sequence does not yield any matches. This aspect of the present invention automatically skips over non-matching text runs within an input string while continuing to perform comparisons for matching runs, resulting in a comparison process that accepts partially matching input sequences.

In a third aspect of the present method, a novel way of resolving conflicts among characters having the same encodings is devised. Occasionally, more than one Chinese character are composed of the same exact components, the construction differing only in the relative placement of the components. To resolve these ambiguous encodings, an additional letter with a prescribed semantic of positional description is appended to each conflicting encoding. Fig. 2 contains an example illustrating this novel technique.

In a fourth aspect of the present method, a novel way of selecting characters matched by the input method is devised. Whenever more than one candidate character matches a user given letter sequence, the candidates are presented to the user for a manual selection. In the prior art, a number is sometimes used as a means of specifying the user choice. While a number is obvious in its meaning since a linear list of candidates are offered up for selection, the present invention chooses to use an alphabetic letter instead. Thus, the letter 'a' signifies choosing the first candidate, 'b' the second, and so forth. The use of an alphabetic letter instead of a number is non-obvious and has never been done in the prior art, as it is not always possible for any given input method since the alphabetic letters are used for encoding Chinese characters and may confuse the system if also used as candidate selection keys. This aspect of the present invention is significant in that it allows the user to keep his fingers on the basal touch typing position (as opposed to having to move them away to type a number), resulting in faster typing speed.

In a fifth aspect of the present method, a novel way of attaching additional information to an input string is devised. Since the present invention only employs the 26 lower case alphabetic letters in constructing input sequences, letters outside of the employed set can be and are used as carriers of additional information about the input sequence. For example, the input sequence "abc6-9" is interpreted to mean 'match all characters defined by the encoding "abc" and with a stroke count of 6 to 9'. Another example is any input sequence beginning with an uppercase letter is

defined to mean "pass through", which means the given input sequence is made the output without interpretation, creating an efficient way of entering English sentences in the midst of Chinese characters.